"BLUP" Estimation in Hierarchical Models

by

Charles E. McCulloch [1]
Biometrics Unit
and
Department of Statistical Science
Cornell University
Ithaca, NY 14853

and

Ziding Feng [2]
Fred Hutchinson Cancer Research Center
1124 Columbia St.
Seattle, WA 98104

BU-1362-MB September 1998

ABSTRACT

Calculation of maximum likelihood estimates for hierarchical models is computationally difficult and a number of other estimation techniques have been studied. Many of the alternatives are based on the related ideas of best linear unbiased prediction (BLUP), maximization of a joint likelihood of the data and unobserved random effects, and penalized quasi-likelihood. We compare and contrast these competing methods with maximum likelihood, point out a number of drawbacks to the competing techniques and suggest reasons for their poor performance.

Keywords: Best linear unbiased prediction, penalized quasi-likelihood, joint maximization, extended quasi-likelihood.

[1] Charles E. McCulloch is Professor of Biological Statistics in the Biometrics Unit of the Department of Statistical Science , Cornell University, Ithaca, NY 14853. Reprint requests should be addressed to him (ceml@cornell.edu).

[2] Ziding Feng is Associate Member of the Fred Hutchinson Cancer Research Center, 1124 Columbia St., Seattle, WA 98104.

1 INTRODUCTION

Hierarchical models are used to accommodate non-normally distributed, correlated data. As such, they are of wide applicability and practical importance (e.g., Breslow and Clayton, 1993). While maximum likelihood and variants are standard for both linear mixed models (e.g., REML) and generalized linear models (e.g., logistic regression), it can be computationally troublesome in hierarchical models due to the need to numerically evaluate high dimensional integrals.

Several approaches have been suggested to avoid these computational problems. McCulloch (1994, 1996, 1997) describes simulation-based approaches to maximum likelihood which are computationally intensive but capable of handling a wide variety of fixed and random effects structures. "Joint-maximization" algorithms have been proposed by a number of authors (Gilmour, Anderson and Rae, 1984; Harville and Mee, 1984; Schall, 1991; Lee and Nelder, 1996). These, following the lead of Henderson *et al.* (1959), arise from maximizing the joint distribution of the observed data and random effects with respect to the parameters *and* the random effects. The earlier papers (Gilmour *et al.*, 1984; Harville and Mee, 1984; Schall, 1991) use linear approximations to arrive at approximate versions of Henderson's Mixed Model Equations (MMEs). The MMEs are particularly nice for the normal-normal linear model since they simultaneously give the MLEs of the fixed effects parameters and the best linear unbiased predictors (BLUPs) of the random effects, while still being computationally efficient. This has led to other suggestions to develop techniques based on BLUPs for GLMMs (McGilchrist, 1994; McGilchrist and Yau, 1995; Engel and Keen, 1994).

Others (Breslow and Clayton, 1993; Wolfinger, 1994) have arrived at essentially the same computational algorithms via different justifications. Breslow and Clayton's approach is that of penalized quasi-likelihood (PQL) in which a penalty function is added to the quasi-likelihood. PQL corresponds to joint maximization with an assumed normal distribution for the random effects.

In Section 2 we describe the joint maximization method and argue that the usual asymptotics for random effects models has the number of random effects increasing. We show by means of analytic calculation that these methods can lead to inconsistent estimates and fail to be invariant in ways they should. Further, we give an example where consistency does not hold even as the sample size increases with fixed numbers of random effects. In Section 3 we show how the methods relate to ML estimation and use it as a guide to understand the performance of the methods. Section 4 offers discussion and conclusion.

2 METHODS OF ESTIMATION

2.1 Henderson's Mixed Model Equations

We begin by first describing the approach Henderson proposed in 1959 which works spectacularly well for the normal-normal model. Henderson studied the model

$$\begin{aligned} \mathbf{Y}|\mathbf{u} &\sim N(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}, \mathbf{R}) \\ \mathbf{u} &\sim N(\mathbf{0}, \mathbf{D}). \end{aligned} \tag{1}$$

He formed the joint distribution, $f_{Y,u}(\mathbf{Y}, \mathbf{u}|\boldsymbol{\beta}, \mathbf{D}, \mathbf{R})$, and observed that maximizing it jointly with respect to $\boldsymbol{\beta}$ and $\mathbf{u}$ gave rise to the following set of equations, the MMEs:

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{D}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \tilde{\boldsymbol{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{Y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Y} \end{bmatrix}. \tag{2}$$

The MMEs are attractive because their solution, $(\hat{\boldsymbol{\beta}}', \tilde{\mathbf{u}}')'$, gives the MLE for $\boldsymbol{\beta}$ and the BLUP of $\mathbf{u}$, which is also the empirical Bayes estimator of $\mathbf{u}$. Coupled with estimation schemes for $\mathbf{D}$ and $\mathbf{R}$, they represent a compact and relatively efficient computational method. For the normal-normal linear model, it can be shown (Speed, 1991) that the REML method (Searle, Casella, and McCulloch, 1992) of estimation of $\mathbf{D}$ corresponds to equating the calculated and expected values of a quadratic form in the BLUP of $\mathbf{u}$ and so is natural to use with the MMEs.

2.2 Methods of Estimation for GLMMs

A logical extension is to try this "joint-maximization" idea for non-normal and/or nonlinear models. And this is exactly what has been proposed by a number of authors. Gilmour, Anderson and Rae (1984) and Harville and Mee (1984) write down the joint maximization equations for binary data with a probit link. The resulting equations are nonlinear, so they linearize them and arrive at approximate analogs of (2). Schall (1991) linearizes the link function directly to derive approximate MMEs for a broad class of GLMMs and relates his work to Stiratelli, Laird and Ware (1984). Recently, Lee and Nelder (1996), have proposed directly maximizing the nonlinear equations rather than making a linear approximation.

Engel and Keen (1994), McGilchrist (1994), and McGilchrist and Yau (1996) make similar proposals within the class of GLMMs based on using BLUPs or approximate BLUPs. Breslow and Clayton (1993) and Wolfinger (1994) derive nearly identical approximate MMEs from alternate justifications (approximate penalized quasi-likelihood and Laplace approximations, respectively). Computationally, PQL is the same as joint maximization with an assumption of normally distributed random effects.

2.3 A Normal-Exponential Example

We now consider a simple example and show that the above methods have two serious drawbacks: inconsistency and lack of invariance. The model we consider is similar to a normal linear mixed model but with non-normally distributed random effects:

$$\begin{aligned} \mathbf{Y}|\mathbf{u} &\sim N(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}^{1/2}, \mathbf{I}) \qquad (3)\\ u_j &\sim \textit{i.i.d.}\ \text{Exponential}(\lambda);\ j = 1, 2, \cdots, q;\ \lambda \text{ known.} \end{aligned}$$

This is a mixed model with linear fixed effects but with the random effects entering in a nonlinear fashion. The data are conditionally independent and normal with known variance 1 and with i.i.d. random effects whose square follows an exponential distribution with known parameter λ. The goal is to estimate $\boldsymbol{\beta}$. The reason for writing the vector of random effects with a square root will become apparent below. The log of the joint distribution of $\mathbf{Y}$ and $\mathbf{u}$ is given by

$$\begin{aligned} ln\, f_{Y,u} &= \text{const} - \frac{1}{2}\left(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}^{1/2}\right)'\left(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}^{1/2}\right) \qquad (4)\\ &\quad + q\, ln(\lambda) - \sum_j \lambda u_j, \end{aligned}$$

and if we set the derivatives of this equal to zero we obtain:

$$\frac{\partial\, ln\, f_{Y,u}}{\partial \boldsymbol{\beta}} = \mathbf{X}'\left(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}^{1/2}\right) = \mathbf{0} \qquad (5a)$$

$$\frac{\partial\, ln\, f_{Y,u}}{\partial \mathbf{u}} = \mathbf{F}^{-1}\mathbf{Z}'\left(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}^{1/2}\right)/2 - \lambda\mathbf{1} = \mathbf{0} \qquad (5b)$$

where $\mathbf{F} = \text{diag}\left\{u_j^{1/2}\right\}$ and note that $\mathbf{F1} = \mathbf{u}^{1/2}$. Working with (5b), we have

$$\mathbf{Z}'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) - \mathbf{Z}'\mathbf{Z}\tilde{\mathbf{u}}^{1/2} = 2\lambda\, \tilde{\mathbf{u}}^{1/2},$$

which is equivalent to

$$\mathbf{Z}'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = (\mathbf{Z}'\mathbf{Z} + 2\lambda\mathbf{I})\, \tilde{\mathbf{u}}^{1/2},$$

or

$$\tilde{\mathbf{u}}^{1/2} = (\mathbf{Z}'\mathbf{Z} + 2\lambda\,\mathbf{I})^{-1}\mathbf{Z}'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}). \tag{6}$$

Plugging this into (5a) gives

$$\mathbf{X}'\mathbf{Y} - \mathbf{X}'\mathbf{X}\tilde{\boldsymbol{\beta}} - \mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z} + 2\lambda\,\mathbf{I})^{-1}\mathbf{Z}'(\mathbf{Y} - \mathbf{X}\tilde{\boldsymbol{\beta}}) = \mathbf{0},$$

and upon setting $\mathbf{P}_\lambda = \mathbf{I} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z} + 2\lambda\,\mathbf{I})^{-1}\mathbf{Z}'$ the above becomes

$$\mathbf{X}'\mathbf{P}_\lambda\mathbf{Y} = \mathbf{X}'\mathbf{P}_\lambda\mathbf{X}\hat{\boldsymbol{\beta}} \qquad \text{or} \qquad \hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{P}_\lambda\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_\lambda\mathbf{Y},$$

provided the inverse exists. It is now straightforward to calculate the expected value of $\hat{\boldsymbol{\beta}}$:

$$\begin{aligned} E[\hat{\boldsymbol{\beta}}] &= (\mathbf{X}'\mathbf{P}_\lambda\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_\lambda(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{E}[\mathbf{u}^{1/2}]) \\ &= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{P}_\lambda\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_\lambda\mathbf{Z}\mathbf{E}[\mathbf{u}^{1/2}]. \end{aligned} \tag{7}$$

Since $E[\mathbf{u}^{1/2}]$ is not zero and, in general, neither is $\mathbf{P}_\lambda\mathbf{Z}, \hat{\boldsymbol{\beta}}$ is biased.

One of the primary reasons for considering random effects models is that there are many levels of the random effect but little information on each one. A random effects model allows sharing of information across the levels. Accordingly, asymptotics for random effects models typically let the number of random effects tend to infinity but do not require the number of observations per random effect to increase (e.g. Westfall, 1986). Under this "usual" scenario, the bias does not tend to zero and the estimator is inconsistent.

As a specific illustration consider the following simple, balanced model with a single random and single fixed effect:

$$\begin{aligned} Y_{ij} &= \beta\, x_i + u_j^{1/2} + \epsilon_{ij}, \quad i = 1, 2, \cdots, n; \quad j = 1, 2, \cdots, q \\ u_j &\sim \textit{i.i.d.}\ \text{Exponential}(\lambda), \\ \epsilon_{ij} &\sim \textit{i.i.d.}\, N(0,1). \end{aligned} \tag{8}$$

Straightforward calculations show that the bias term in (7) for model (8) is given by

$$\frac{\Sigma_i x_i \sqrt{\pi\lambda}}{(\Sigma_i x_i^2)(n+2\lambda) - (\Sigma_i x_i)^2}.$$

This is independent of q, the number of levels of the random factor. So (unless $\Sigma_i x_i = 0$) the bias does not go to zero as q increases, the usual sort of asymptotics for random effects. Worse, if the x_i are all constant (equal to, say, c) then the bias is $\sqrt{\pi/\lambda}/(2c)$ and does not even go to zero as both the levels of the random effect *and* the number of observations per level increase.

We now return to the reason for writing the model as in (3). Clearly, an equivalent way to write (3) would be to let $\mathbf{w} = \mathbf{u}^{1/2}$ and write:

$$\begin{aligned} \mathbf{Y}|\mathbf{u} &\sim N(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{w}, \mathbf{I}) \\ w_i &\sim i.i.d.\ f_w(w|\lambda), \lambda \text{ known}, \end{aligned} \tag{9}$$

where f_w is the density of a random variable whose square is exponential. The joint maximization equations under this formulation are:

$$\frac{\partial\, ln\, f_{Y,u}}{\partial \boldsymbol{\beta}} = \mathbf{X}'\left(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{w}\right) = \mathbf{0} \tag{10a}$$

$$\frac{\partial\, ln\, f_{Y,u}}{\partial \mathbf{u}} = \mathbf{Z}'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{w}) - 2\lambda\mathbf{w} + \mathbf{1}/\mathbf{w} = \mathbf{0} \tag{10b}$$

where $\mathbf{1}/\mathbf{w}$ is interpreted as a column vector with elements $1/w_j$. Equation (10a) is the same as (5a) but the equation for $\mathbf{w} = \mathbf{u}^{1/2}$ is different. Hence the solutions for the random effects are not the same which leads to different solutions for $\boldsymbol{\beta}$. This simple example shows that joint maximization solutions are not invariant to writing the model in a different but statistically completely equivalent way! So we cannot count on joint maximization to lead to sensible estimators.

Do these two parameterizations give results which are similar? Unfortunately no, as the following example shows. We simulated model (8) with $\beta = 5$, $n = 2$, $x_i = i/n$, $\lambda = 1$ and various values of q. We also calculated the maximum likelihood estimator of β. Simulation details are given in the Appendix. Table 1 shows the results.

The two parameterizations of the model give very different average estimates. The estimated bias for the $\mathbf{Zu}^{\frac{1}{2}}$ parameterization is very close to the calculated value from (7) of $6\sqrt{\pi}/11 \doteq 0.967$. The ML estimator does not exhibit any bias.

Table 1: Average values of estimates of β from paired data from the normal/exponential model using various estimation methods. Standard errors are in parentheses. The true value of β is 5.

	Joint maximization		ML estimate
Sample Size	Using $\mathbf{Zu}^{\frac{1}{2}}$	Using $\mathbf{Zw}$	
20	5.979(.007)	5.163(.007)	5.017(.007)
50	5.968(.006)	5.150(.005)	5.004(.005)
100	5.958(.007)	5.139(.006)	4.993(.006)
200	5.977(.007)	5.157(.007)	5.011(.007)

3 BLUP ESTIMATION AND THE ML EQUATIONS

Whenever the marginal density of $\mathbf{Y}$ is formed as a mixture as in (1), with separate parameters for $f_{Y|u}$ and f_u, then the ML equations for $\boldsymbol{\beta}$ and $\mathbf{D}$ take the following form:

$$E\left[\frac{\partial\, ln\, f_{Y|u}(\mathbf{Y}|\mathbf{U},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}|\mathbf{Y}\right] = \mathbf{0} \tag{11a}$$

$$E\left[\frac{\partial\, ln\, f_u(\mathbf{U}|\mathbf{D})}{\partial\mathbf{D}}|\mathbf{Y}\right] = \mathbf{0}. \tag{11b}$$

To see how these relate to BLUP, it is instructive to start with the linear mixed model:

$$\begin{aligned} \mathbf{Y}|\mathbf{u} &\sim N(\mathbf{X}\boldsymbol{\beta} + \mathbf{Zu}, \mathbf{I}\sigma^2) \\ \mathbf{u} &\sim N(\mathbf{0}, \mathbf{D}). \end{aligned} \tag{14}$$

For this model, equation (11a) for $\boldsymbol{\beta}$ is

$$E\left[\mathbf{X}'(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u})/\sigma^2|\mathbf{Y}\right]=\mathbf{0}$$

or, since $\tilde{\mathbf{u}}=E[\mathbf{u}|\mathbf{Y}]$ (Searle, Casella and McCulloch, 1992),

$$\mathbf{X}'\mathbf{Y}-\mathbf{X}'\mathbf{X}\boldsymbol{\beta}-\mathbf{X}'\mathbf{Z}\tilde{\mathbf{u}}=\mathbf{0},$$

which is the system of equations for $\boldsymbol{\beta}$ from (2), the MMEs. To solve this for $\boldsymbol{\beta}$ we need $\tilde{\mathbf{u}}=E[\mathbf{u}|\mathbf{Y}]$. Since $ln\, f_{Y,u}=ln\, f_{u|Y}+ln\, f_Y$ and since f_Y does not involve $\mathbf{u}$, $\partial\, ln\, f_{Y,u}/\partial\mathbf{u}=\partial\, ln\, f_{u|Y}/\partial\mathbf{u}$. Setting this derivative equal to zero gives the mode of $f_{u|Y}$. Since the distribution of $\mathbf{u}$ given $\mathbf{Y}$ is normal, the mode is the mean which is $E[\mathbf{u}|\mathbf{Y}]=\tilde{\mathbf{u}}$. Thus the joint maximization equation for $\mathbf{u}$ finds $E[\mathbf{u}|\mathbf{Y}]$ which is needed to solve for the MLE of $\boldsymbol{\beta}$. This is why joint maximization works for the normal-normal linear model.

In contrast, for non-normal distributions, the conditional distribution of $\mathbf{u}$ given $\mathbf{Y}$ is not normal and hence maximizing $f_{Y,u}$ with respect to $\mathbf{u}$ locates the mode which may not be $E[\mathbf{u}|\mathbf{Y}]=\tilde{\mathbf{u}}$. Furthermore, the likelihood equation (11a) may involve functions more complicated than $\mathbf{u}$ alone. Hence other conditional expected values of functions of $\mathbf{u}$ are required to calculate the MLE of $\boldsymbol{\beta}$.

For example, for the model given by (3), the ML equation for $\boldsymbol{\beta}$ is

$$\mathbf{X}'\mathbf{Y}-\mathbf{X}'\mathbf{X}\boldsymbol{\beta}-\mathbf{X}'\mathbf{Z}\,E[\mathbf{u}^{1/2}|\mathbf{Y}]=\mathbf{0}. \tag{15}$$

So what is required is $E[\mathbf{u}^{1/2}|\mathbf{Y}]$ ($\neq E[\mathbf{u}|\mathbf{Y}]^{1/2}=\tilde{\mathbf{u}}^{1/2}$), not $\tilde{\mathbf{u}}=E[\mathbf{u}|\mathbf{Y}]$. The joint maximization equations solve for the mode of $f_{u|Y}$:

$$\tilde{m}(\mathbf{Y})=\max_u f_{u|Y},$$

and then $\tilde{m}(\mathbf{Y})^{1/2}$ is then used in (12) to solve for $\boldsymbol{\beta}$. Engel and Keen (1996), in their discussion of the Lee and Nelder paper, note this discrepancy but seem to downplay its importance. These two quantities will not usually be equal unless var($\mathbf{u}|\mathbf{Y}$) is zero. This is tantamount to there being an infinite amount of information in $\mathbf{Y}$ about $\mathbf{u}$. We have argued in Section 2.3 that this will not be the usual case in random effects models.

In summary, in the linear mixed model, equation (11a) for $\boldsymbol{\beta}$ is linear in $\mathbf{u}$. Hence the only conditional expectation needed is $E[\mathbf{u}|\mathbf{y}]$, which is found exactly by solving $\partial \, ln \, f_{Y,u}/\partial \mathbf{u} = \mathbf{0}$, i.e., one of the joint maximization equations. For hierarchical models in general, not only does solving $\partial \, ln \, f_{Y,u}/\partial \mathbf{u} = \mathbf{0}$ not give $E[\mathbf{u}|\mathbf{Y}]$ but, even if we could easily calculate $E[\mathbf{u}|\mathbf{Y}]$, it is not necessarily the needed ingredient to use in solving (11a).

4 DISCUSSION AND CONCLUSIONS

The example in Section 2.3 and the discussion in 2.4 show that joint maximization approaches cannot be reliably depended upon to yield good estimators. The warning not to try to estimate "missing data" along with parameters was noted some time ago by Little and Rubin (1983). We have been able to show explicitly the possible consequences of trying to do so for a class of nonlinear mixed models. To our knowledge, only simulation results have been available previously. In contrast, maximum likelihood avoids the problems noted here. For example, the ML equations for $\boldsymbol{\beta}$ under model (3) are given by (13):

$$\mathbf{X}'\mathbf{Y} - \mathbf{X}'\mathbf{X}\boldsymbol{\beta} - \mathbf{X}'\mathbf{Z}\, E[\mathbf{u}^{1/2}|\mathbf{Y}] = \mathbf{0},$$

while the equations under the equivalent model, (9), are given by:

$$\mathbf{X}'\mathbf{Y} - \mathbf{X}'\mathbf{X}\boldsymbol{\beta} - \mathbf{X}'\mathbf{Z}\, E[\mathbf{w}|\mathbf{Y}] = \mathbf{0}.$$

These are equivalent equations to solve since $\mathbf{u}^{1/2} = \mathbf{w}$ and we can appeal to basic properties of expectation. ML estimates are also consistent under the appropriate regularity conditions. For example, ML estimation would generate consistent estimates of $\boldsymbol{\beta}$ for the model given by (8) for fixed n, but increasing q, unlike joint maximization.

It has been noted (e.g., Harville and Mee, 1984; Firth, 1996) that the joint maximization equations are computationally equivalent to the Bayesian modal posterior estimates of $\boldsymbol{\beta}$ and the random effects if one adopts a flat prior for $\boldsymbol{\beta}$. So this type of Bayesian analysis will suffer the same shortcomings as joint maximization. One way to view the lack of invariance with joint maximization is the simple lack of invariance of a mode under transformation of a variable. For example, an exponentially distributed random variable has a mode of zero, while the square of an exponential has a mode at $1/\sqrt{2\lambda}$. Even if this is "back-transformed" it gives a different answer. Thus the problems can be traced to using modes instead of expectations. This lack of invariance is a known drawback in using a Bayesian analysis with modal posterior estimates but was not mentioned or documented as a criticism by Firth (1996).

5 REFERENCES

Aptech Systems (1990), *GAUSS 2.1 User's Manual*, Kent, WA: Author.

Breslow, N.E. and Clayton, D.G. (1993), "Approximate Inference in Generalized Linear Mixed Models," *Journal of the American Statistical Association* **88**, 9-25.

Engel, B. and Keen, A. (1994), "A Simple Approach for the Analysis of Generalized Linear Mixed Models," *Statistica Neerlandicá* **48**, 1-22.

Engel, B. and Keen, A. (1996), "Comments on Lee and Nelder: 'Hierarchical Generalized Linear Models,'" *Journal of the Royal Statistical Society, Series B*, **58**, 656-7.

Firth, D. (1996), "Comments on Lee and Nelder: 'Hierarchical Generalized Linear Models' ," *Journal of the Royal Statistical Society, Series B* **58**, 661.

Gilmour, A.R., Anderson, R.D., and Rae, A.L. (1985), "The Analysis of Binomial Data by a Generalized Mixed Model," *Biometrika* **72**, 593-599.

Harville, D.A. and Mee, R.W. (1984), "A Mixed Model Procedure for Analyzing Ordered Categorical Data," *Biometrics* **40**, 393-408.

Henderson, C.R., Kempthorne, O., Searle, S.R., and VonKrosigk, C.N. (1959), "Estimation of Environmental and Genetic Trends from Records Subject to Culling," *Biometrics* **30**, 583-588.

Lee, Y. and Nelder, J.A. (1996), "Hierarchical Generalized Linear Models," *Journal of the Royal Statistical Society, Series B* **58**, 619-678.

Little, R.A. and Rubin, D.B. (1983), "On Jointly Estimating Parameters and Missing Data by Maximizing the Complete Data Likelihood," *The American Statistician* **37**, 218-220.

McCulloch, C.E. (1994), "Maximum Likelihood Variance Components Estimation for Binary Data," *Journal of the American Statistical Association* **89**, 330-335.

McCulloch, C.E. (1996), "Maximum Likelihood Estimation for Serially Correlated Longitudinal Data," Biometrics Unit Technical Report BU-1337-M, Cornell University, Ithaca, NY.

McCulloch, C.E. (1997), "Maximum Likelihood Algorithms for Generalized Linear Mixed Models." *Journal of the American Statistical Association***92**, 162-170.

McGilchrist, C.A. (1994), "Estimation in Generalized Mixed Models," *Journal of the Royal Statistical Society, Series B* **56**, 61-69.

McGilchrist, C.A. and Yau, (1995), "The Derivation of BLUP, ML, and REML Estimation Methods for Generalized Linear Mixed Models," *Communications in Statistics, Theory and Method* **24**, 2963-2980.

Robinson, G.K. (1991), "That BLUP Is a Good Thing: Estimation of Random Effects," *Statistical Science* **6**, 15-31.

Schall, R. (1991), "Estimation in Generalized Linear Models with Random Effects," *Biometrika* **78**, 719-727.

Searle, S.R., Casella, G., and McCulloch, C.E. (1992), *Variance Components*, New York: Wiley.

Speed, T. (1991), "Comments on Robinson: Estimation of Random Effects," *Statistical Science* **6**, 42-44.

Stiratelli, R., Laird, N., and Ware, J.H. (1984), "Random Effects Models for Serial Observations with Binary Response," *Biometrics* **40**, 961-971.

Westfall, P.H. (1986), "Asymptotic Normality of the ANOVA Estimates of Components of Variance in the Nonnormal, Unbalanced, Hierarchical Mixed Model," *Annals of Statistics* **14**, 1572-1582.

Wolfinger, R. (1994), "LaPlace's Approximation for Nonlinear Mixed Models," *Biometrika* **80**, 791-795.

Appendix

The simulation was run using GAUSS (Aptech Systems, 1990). Data sets were generated using the built in uniform (to generate exponential variates) and normal random number generators. Estimates under model (8) were calculated using the formula for $\hat{\boldsymbol{\beta}}$. Estimates under model (9) were formed using a substitution algorithm to solve (10a) and (10b).

To calculate the maximum likelihood estimators we first form the log likelihood as

$$\begin{aligned} ln\, L &= -\frac{nq}{2}\, ln(2\pi) + q\, ln(2\lambda) - \frac{1}{2}\sum_{i,j} \tilde{Y}_{ij}^2 \\ &+ \sum_{j=1}^{q} ln\left[\frac{1}{n+2\lambda} - \frac{e^{\Sigma \tilde{Y}_{ij}^2/(4\lambda+2n)}}{(n+2\lambda)^{3/2}} \sum \tilde{Y}_{ij}\sqrt{2\pi}\,\Phi\left(\frac{\Sigma \tilde{Y}_{ij}}{(n+2\lambda)^{1/2}}\right)\right], \end{aligned}$$

where $\tilde{Y}_{ij} = Y_{ij} - \beta X_{ij}$. We then numerically maximized $ln\, L$ using the OPTMUM routine in GAUSS. The number of replications for the simulation was chosen to achieve standard errors of .007 or less.